Weekly report

1 Done

1.1 Vast Modification

The red part is not done yet.

Summary Review

The summary review listed seven main issues with the initial submission, listed below with our responses.

1. **Better describe motivation, design process, and unique selling point of the system. [R2,R3,R4]**

   As mentioned by Reviewer 2, we added application scenario (providing privacy preservation before sharing data) as motivation, as well as selling point of the system.

   We designed interfaces based on the task requirements (Section 3). Now, we mentioned that at the beginning of Section 5. Moreover, we constructed the links between views and specific task requirements when describing visual designs. We think this improvement can facilitate readers in understanding our design process.

2. **Give motivation behind individual design choices. [R3,R4]**

   On this issue, Reviewer 3 raised a question about the choice of common metrics for the hub fingerprint. Actually, the three metrics are selected by users in the priority view. We consider that they play an important role to the graph. So, our system automatically provides these metrics. This reason is now explained in the second paragraph about the hub fingerprint protector.

   On the other hand, Reviewer 4 raised a series of question about our algorithm. need check finally The reason why we achieved privacy preservation only by adding edges and not remove them is that employing both schemes simultaneously may cause conflict. For example, the later processing may remove the edges added in the previous processing. Besides, we choose adding edges rather than removing edges, because adding edges helps protect characteristics, like high degree. We have provided details in the Section 4.2.

3. **Improve the reporting of the evaluation. [R2,R3]**

   We have reorganized both Section 6 (case studies) and Section 7.1. We introduced more details on the results of two case studies. Simultaneously, we spent more space to convey the evaluation of our system by experts.

4. **Assess and discuss the limitations of the tool in an honest and complete manner (low-level vs. high-level utility [R1], false sense of privacy [R4], scalability and usability issues [R5] etc.).**

   We have added high-level utility evaluation to our system. Related information can be found in the second part (Other Changes).
   As mentioned by Reviewer 4, our approach can remove the three privacy issues mentioned in the paper. However, the processed data may face other threats from other issues. Therefore, we defined the target users as the individuals who are knowledgeable about anonymization approaches. Moreover, we listed extending protectors as future work.
   Other issues are discussed in Section 7. For scalability, we provided detailed discussion from the perspectives of both visual expression and algorithms.

5. **State clearly that the tool is a first attempt to establishing very basic privacy protection in a visual manner. It does not provide full anonymity and using knowledge about another graph with the same user set, the users of an anonymized dataset can likely be re-identified. [R4].**

   In the section 8, we explained the limitation of our system from two aspects: resisting partial attacks and focusing on only simple graphs. We also consider them in our future work.

6. **Provide more information on the utility aspects [R1,R5]**

   We added graph clustering to show utility changes. Moreover, we introduced more details about utility in Section 6.

7. **Revise the writing of the paper for a more precise wording (semi-run, scrubbing, etc.), a clear structure (Sec.2, Sec.5.2,…) , and short inline explanations/definitions that clarify terms on the spot (k-anonymity, utility,…)**

   We did our best to improve writing. Thanks reviewers for your patience to pointing out our typos.

Other Changes

Regarding more minor comments and issues, we have made the following edits:

1. We simplified Figure 6 for saving space.
2. The Figure 12(a) is removed.

3. In Figure 13(a), we added details about the amount changes of low-degree nodes .

Responses

Reviewer 1:

1. **Regarding the limited notion of Utility.**

   Now, users can check graph clustering to evaluate utility changes.

2. **Regarding the writing style of the paper.**

   We'd like first to thank you for the detailed reviews. As your advices, we made a series of modifications. In addition, we want to explain that the "closeness" in *t-closeness* refers to different meaning from the "closeness" in *closeness centrality*. To avoid misunderstanding, we removed the mention of *t-closeness* in the introduction.

Reviewer 2:

1. **Anonymisation is also discussed in the following article, but it is accomplished through aggregation.**

The related paper is added to the subsection titled "Privacy-aware Visualizations".

Reviewer 3:

1. **A comparison of the results of using GraphProtector with the results of other comparable systems would be useful. It does not need to be extremely detailed, but should make it clearer what GraphProtector offers that comparable systems don't. They mentioned that the reviewers have issues observing tabular data, but it is not clear that this is the only alternative.**

   As far as we know, the most comparable system is SecGraph, which allows users to call a variety of automatic algorithms for both privacy preservation and result evaluation. However, SecGraph has no ability in integrating privacy-preserving algorithms. Furthermore, it provides no visual expression but results in the form of tables. Thus, we only discussed the differences between automatic algorithms and our approach based on expert reviews in Section 7.1.

Reviewer 4:

1. **The assumptions about the threat model are unclear. For example, is the goal to provide privacy for a set of users or all the users? I assume, even if k-**

anonymity is provided for some users, still some users not in the k-anonymity group can be re-identified.

We detect various characteristics among the entire graph rather than a set of users. In the degree protector and hub fingerprint protector, our system count numbers for all characteristics appearing in the graph. Therefore, all users are take in to consideration. However, subgraph is so complicated that we can hardly enumerate all of them. Based on the assumption of potential attacks, analysts can take targeted defenses by defining specific subgraphs. After processing, the users involved in related subgraphs will under protection.

2. **Also, the provided protection mechanisms do not protect against all the de-anonymization attacks. For example, it has been shown that if the adversary have knowledge about another graph with the same user set then even adding/ removing edges to the anonymized graph might not be effective.**

   K-anonymity approaches need knowledge about potential attacks. In the case mentioned above, analysts can apply subgraph protector and load the subgraph constructed by the user sets exposed to attackers.

3. **The subgraph protector has been quite simplified. Even if k subgraphs exist in the network, they are part of larger subgraphs that still can re-identify them. I assume the user needs to tries k for subgraphs with different sizes. However, how he can decide which one is better? It seems the only metric that is provided for measuring privacy is K.**

   K-anonymity approaches need users to have knowledge about the potential risks. We allow users to detect subgraph issues by defining classical structures and loading a subgraph. Thus, they can protect any subgraphs they care.

4. **It is assumed that the user of the tool has a great knowledge of privacy threats and the required solutions for them. However, if the user is not competent, then just employing this tool randomly can be even more dangerous because it gives a false sense of privacy protection.**

   We updated our definition of target users as someone who work with sensitive network datasets, require flexible and personalized privacy preservation, and are knowledgeable about anonymization approaches for graphs.

5. **How the researcher can choose some nodes in this large network to for example prioritize or lock them? Does the tool provide a search function based on some non-graph attributes, such as users' names, etc?**

In the priority view, nodes are selected according to the metric values. Users can realize constructs, like ``prioritize the 30% of nodes with the lowest degree," or ``if possible, do not touch the 1% of nodes with the highest betweenness."

6. **More explanation should be provided for using multiple protectors. What is the actual algorithm for this? How combining techniques impact privacy and utility? I assume after for example, employing the hub fingerprint protector, the degree distribution would change and again degree protector should be applied.**

Actually, the algorithms can inspect if new issues are generated based on the previous user settings.

Reviewer 5:

1. **Related Work**

We have restructured the subsections here per R5's suggestion and edited them so that the flow and organization is improved. We have also added a reference for general graph visualization techniques.

2. **Task Requriements**

- What does utility mean? -> short definition

We refer utility as the similarity of structural properties. Now, we explained utility at the beginning of the section.

- TR1: while I like the addition of statistical plots, why start with a simple node-link visualization in the first place? -> reason

The reason is that when a number of statistical plots can be provided, users need to observe the overview of the graphs to select the significant ones. We gave this reason in the Section 4.3.

- TR2: what may be the user's needs that are so important that they should interfere with privacy preservation -> give an example to make this clearer

Actually, setting priority exert few effects on privacy preservation. By setting priority, users can guide automatic algorithms to better choice, which can also achieve privacy preserving requirements.

- TR3: what are costs in this context -> short definition

We changed “costs” to “utility costs”. Combined with the definition of utility, we think readers can understand it.

3. Visualization

The Utility View: Why only provide average values

We calculated the modification of vectors by similarity measure rather than average. We also provide four common metrics for similarity: Euclidean distance, Manhattan distance,cosine similarity and Jaccard similarity. They are employed by the experts in graph privacy as well.

The Graph Protector View

We disabled the privacy preservation with the highest degree to simulate a situation. Supposed that the three high-degree nodes are public people, like government officers or stars. Due to the need of work, they have to show their identities to everyone. Therefore, there is no need to anonymize their high-degree features and cause more utility loss. Note that to protect the privacy of the people who have relationship with them, like their family or friends, we choose them as hubs in the hub finger protector.

4. Discussion

Automation

1.2 Group Meeting

Presentation and blog writing.

1.3 Interview the Interns

2 Progress

Item	Deadline	Current progress	Remark
Vast modification	6.27	More than half writing improvements are done.	
T-ITS modification	9.15	Summarize reviews.	
Courseware revision	9.1	-	
Privacy program	10.31	Surveying.	Pausing.